Verbal vs. Mathematical Aptitude: An Investigation on the SAT's Predictability

This study seeks to answer a compelling question: Is a higher verbal or mathematical aptitude, as evidenced by SAT scores, correlated with greater success in college, as measured by first-year GPA? This question delves into the heart of academic achievement and aims to unravel whether specific aptitudes, namely in verbal or mathematical domains, are significant predictors of collegiate success. We analyze data from OpenIntro, originally sourced by the Educational Testing Service, focusing on the "SAT and GPA data" dataset. This dataset includes 1000 students from an undisclosed college, providing a unique opportunity to examine how different aptitudes, as reflected in standardized test scores, relate to academic performance in the first year of college. The hypothesis driving this inquiry is that students with higher aptitude in a particular domain (verbal or mathematical) might demonstrate differing levels of success in their initial college year, potentially reflecting the varying demands and skills required in different academic disciplines.

The dataset used for this study is sourced from the Educational Testing Service on OpenIntro. I took the "SAT and GPA data" in order to find an answer to my question. Each student's record consists of six variables: gender(sex), verbal SAT percentile (sat_v), math SAT percentile (sat_m), combined SAT percentile (sat_sum), high school GPA (hs_gpa), and first-year college GPA (fy_gpa). The SAT percentile scores range from 0 to 100 for each section, leading to a combined total ranging from 0 to 200. Both high school and college GPAs are unweighted, spanning from 0.00 to 4.00. The diverse scales and units of measurement present in the dataset necessitated the application of logarithmic transformations for a more accurate analysis of the causal relationships.

In our initial Ordinary Least Squares (OLS) regression model, we examined the relationship between high school GPA, a proxy for academic effort, and first-year college GPA. Our analysis found a significant correlation: without controlling for other variables, a 1% increase in high school GPA led to a 0.96% rise in college GPA[1]. Introducing the combined SAT score percentiles as a control for innate ability slightly diminished this correlation, with a 1% increase in high school GPA then correlating with a 0.75% increase in college GPA[2]. Conversely, focusing on the combined SAT percentile scores revealed a more pronounced initial effect on college GPA. Without controlling for high school GPA, a 1% increase in SAT scores was associated with a 1.02% increase in college GPA[3]. However, this effect was reduced to a 0.62% increase in college GPA when high school GPA was included as a control[4]. Therefore, we can see that ability in this case accounts for an OVB of 0.21.

To further explore the interplay between effort and aptitude, we conducted two separate regression analyses. The first targeted students with higher verbal than mathematical SAT scores, and the second focused on those with higher scores in math. Both models controlled for the combined value of SAT section percentiles to adjust for overall ability. The findings revealed that for students with a verbal aptitude, a 1% increase in high school GPA correlated with a 0.74% increase in college GPA, a statistically significant effect[5]. Similarly, for students with a mathematical aptitude, a 1% increase in high school GPA was associated with a 0.77% increase in college GPA[6]. These results suggest that while specific aptitudes (verbal or mathematical) slightly modulate the impact of high school effort on college GPA, the differences are not substantial. However, the limited nature of our analysis prevents us from making definitive conclusions about the magnitude of these differences, highlighting the need for more controlled studies, such as a randomized controlled trial, to gain clearer insights.

The findings of this study suggest that both verbal and mathematical aptitudes, as measured by SAT scores, are correlated with first-year college GPA, alongside the influence of high school GPA. However, the specific impact of each type of aptitude varies slightly, indicating that different skill sets may confer advantages in different academic contexts. In order to have more conclusive results without any biases, the study should consider that there are several factors that can skew SAT score results including: location and availability of testing sites, income of families in order to pay for tutors, and education of the subjects in high school. With these factors in mind, we cannot definitely say accept our hypothesis as the difference in aptitudes is not great enough without these variables accounted for. If we were to conduct a randomized controlled trial(RCT) with enough data to account for these OVBs, we could perhaps see a change in the answer to the hypothesis. Additionally, this study highlights the need for further research to explore the nuanced ways in which different aptitudes contribute to academic success in various disciplines. This study underscores the importance of considering a range of factors, including both general academic effort and specific aptitudes, in predicting and supporting college success.

*ChatGPT was used in order to reform and reword the 5th and 6th paragraphs in order to condense the information into the appropriate length and allow for greater clarity.

[1] *First Regression: College GPA on High School GPA*

Percentage Change in College GPA = α + βPercentage Change in High School GPA

```
shortreg  <- lm(log_fy_gpa ~ log_hs_gpa,
                data = hs_college_gpa_filtered)
summary(shortreg)
```

```
Call:
lm(formula = log_fy_gpa ~ log_hs_gpa, data = hs_college_gpa_filtered)

Residuals:
     Min       1Q   Median       3Q      Max
-1.90447 -0.13200  0.05542  0.19652  0.78685

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.25017    0.06375  -3.924  9.3e-05 ***
log_hs_gpa   0.96126    0.05487  17.518  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.304 on 995 degrees of freedom
Multiple R-squared:  0.2357,    Adjusted R-squared:  0.2349
F-statistic: 306.9 on 1 and 995 DF,  p-value: < 2.2e-16
```

[2] *Second Regression: College GPA on High School GPA with SAT Control*

Percentage Change in College GPA = α + βPercentage Change in High School GPA + γPercentage Change in Summed SAT Percentiles

```
longreg  <- lm(log_fy_gpa ~ log_hs_gpa + log_sat_sum,
               data = hs_college_gpa_filtered)
summary(longreg)
```

```
Call:
lm(formula = log_fy_gpa ~ log_hs_gpa + log_sat_sum, data = hs_college_gpa_filtered)

Residuals:
     Min       1Q   Median       3Q      Max
-1.89586 -0.11288  0.04159  0.18166  0.74059

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.92998    0.31758  -9.226   <2e-16 ***
log_hs_gpa   0.75287    0.05824  12.926   <2e-16 ***
log_sat_sum  0.62934    0.07317   8.601   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2934 on 994 degrees of freedom
Multiple R-squared:  0.2887,    Adjusted R-squared:  0.2872
F-statistic: 201.7 on 2 and 994 DF,  p-value: < 2.2e-16
```

[3] *Third Regression: College GPA on SAT Sum*

Percentage Change in College GPA = α + βPercentage Change in Summed SAT Percentiles

```
shortregsat  <- lm(log_fy_gpa ~ log_sat_sum,
                   data = hs_college_gpa_filtered)
summary(shortregsat)
```

```
Call:
lm(formula = log_fy_gpa ~ log_sat_sum, data = hs_college_gpa_filtered)

Residuals:
     Min       1Q   Median       3Q      Max
-1.87171 -0.15699  0.05548  0.21447  0.63988

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.89012    0.33354  -11.66   <2e-16 ***
log_sat_sum  1.02275    0.07188   14.23   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.317 on 995 degrees of freedom
Multiple R-squared:  0.1691,    Adjusted R-squared:  0.1682
F-statistic: 202.5 on 1 and 995 DF,  p-value: < 2.2e-16
```

[4]*Fourth Regression: College GPA on SAT Sum with High School GPA Control*

Percentage Change in College GPA = α + βPercentage Change in Summed SAT Percentiles + γPercentage Change in High School GPA

```
longregsat  <- lm(log_fy_gpa ~ log_sat_sum + log_hs_gpa,
                  data = hs_college_gpa_filtered)
summary(longregsat)
```

```
Call:
lm(formula = log_fy_gpa ~ log_sat_sum + log_hs_gpa, data = hs_college_gpa_filtered)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.89586 -0.11288  0.04159  0.18166  0.74059 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2.92998    0.31758  -9.226   <2e-16 ***
log_sat_sum  0.62934    0.07317   8.601   <2e-16 ***
log_hs_gpa   0.75287    0.05824  12.926   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2934 on 994 degrees of freedom
Multiple R-squared:  0.2887,	Adjusted R-squared:  0.2872 
F-statistic: 201.7 on 2 and 994 DF,  p-value: < 2.2e-16
```

[5]*Fifth Regression: College GPA on High School GPA with Verbal Aptitude Filter and SAT Control*

Percentage Change in College GPA = α + βPercentage Change in High School GPA + γPercentage Change in Summed SAT Percentiles

- Data = Verbal SAT Percentile > Math SAT Percentile

```
verbal_higher <- filter(hs_college_gpa_filtered, sat_v > sat_m)
model_verbal_higher <- lm(log_fy_gpa ~ log_hs_gpa + log_sat_sum, data = verbal_higher)
summary(model_verbal_higher)
```

```
Call:
lm(formula = log_fy_gpa ~ log_hs_gpa + log_sat_sum, data = verbal_higher)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.33151 -0.13145  0.03308  0.19179  0.57021 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.5733     0.5873  -4.382 1.77e-05 ***
log_hs_gpa    0.7403     0.1128   6.563 3.26e-10 ***
log_sat_sum   0.5565     0.1351   4.118 5.27e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2957 on 238 degrees of freedom
Multiple R-squared:  0.2874,	Adjusted R-squared:  0.2814 
F-statistic:    48 on 2 and 238 DF,  p-value: < 2.2e-16
```

[6] *Sixth Regression: College GPA on High School GPA with Math Aptitude Filter and SAT Control*

Percentage Change in College GPA = α + βPercentage Change in High School GPA + γPercentage Change in Summed SAT Percentiles

- Data = Math SAT Percentile > Verbal SAT Percentile

```
math_higher <- filter(hs_college_gpa_filtered, sat_m > sat_v)
model_math_higher <- lm(log_fy_gpa ~ log_hs_gpa + log_sat_sum, data = math_higher)
summary(model_math_higher)
```

```
Call:
lm(formula = log_fy_gpa ~ log_hs_gpa + log_sat_sum, data = math_higher)

Residuals:
    Min      1Q  Median      3Q     Max 
-1.8925 -0.1040  0.0415  0.1784  0.6106 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.23794    0.38436  -8.424  < 2e-16 ***
log_hs_gpa   0.77109    0.06957  11.083  < 2e-16 ***
log_sat_sum  0.69042    0.08841   7.810 2.08e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2898 on 706 degrees of freedom
Multiple R-squared:  0.3037,	Adjusted R-squared:  0.3018 
F-statistic:   154 on 2 and 706 DF,  p-value: < 2.2e-16
```